Using Student Assessment Engagement as a Measure of Student SEL and School Engagement

April 2017

Jim Soland, Ph.D.
Nate Jensen, Ph.D.

nwea

Measuring What Matters™

Narrative Proposal

Significance of the Proposed Assessment

1. Response-time Effort: A Metric that Uses Achievement Test Metadata to Measure Self-management and Academic Motivation. Metadata that are often captured and discarded when students take achievement tests on a computer can transform processes for identifying, monitoring, and supporting students who might benefit from social-emotional learning (SEL) interventions. In this submission, we present a metric called response-time effort (RTE) that relies on such metadata. The measure uses item response times, or the seconds that elapse between when a question is presented and answered, to identify when students respond to a test question so quickly they could not have understood its content. This behavior is referred to as "rapid guessing" (Wise & Kong, 2005). RTE measures the proportion of items from a test on which a student did not rapidly guess. For example, a student with an RTE of .95 rapidly guessed on 5% of the items. The metric is associated with more than a decade of validity evidence supporting its use as a measure of test-taking engagement, chronicled by Wise (2015). As importantly, RTE metrics are scalable. This fall, RTE will be incorporated into standard reports for any student taking NWEA's Measures of Academic Progress (MAP), an interim assessment suite used to measure mathematics, reading, language usage, and science achievement in more than 6,500 U.S. school systems.

Recent research of ours conducted in collaboration with Santa Ana Unified School District (SAUSD) shows that RTE is useful as much more than a proxy for test motivation. Our study indicates that rapid-guessing behavior is associated with low self-management scores on district-administered SEL surveys (Soland, Jensen, Keys, Bi, & Wolk, 2017).[i] This relationship makes intuitive sense. Self-management can be defined as whether students maintain control over their thoughts, behaviors, and emotions. The construct measures whether students perform a collection of observable behaviors like coming to class prepared, following directions, and working independently.[ii] (Specific survey items used by SAUSD are in Appendix 1.) Generally, students with low self-management have trouble staying focused and completing tasks. One could imagine that a student who struggles with self-management might also have difficulty maintaining focus during a test.

While the ability to complete small tasks may seem trivial, self-management predicts important outcomes like grades and graduation rates. The theory connecting self-management to these outcomes is straightforward, if multi-faceted. Students who lack academic self-efficacy—meaning they do not believe they are capable of completing academic tasks—have little incentive to undertake such tasks. Therefore, self-efficacy is a fundamental building block of student motivation (Bandura, 1997). A lack of academic motivation, in turn, can manifest itself in behaviors like failing to complete coursework and coming to class unprepared. Therefore, self-management might be viewed as a collection of behaviors that are outward signs of low motivation and self-efficacy, and that oftentimes suggest a student is at risk of dropping out. In our research, we show that RTE is associated with other behaviors that are warning signs of low academic motivation including course failures, suspensions/expulsions, and absenteeism. For

1

example, students who rapidly guessed on 10% or more of the items on a test were absent from school an additional day, on average, compared to students who did not rapidly guess.

2. RTE is a Direct Measure of Rapid Guessing, and a Proxy for Low Self-management. As a measure of self-management, RTE has several advantages over student self-report and teacher observation measures. Unlike surveys, RTE directly measures a student behavior—rapid guessing—and does so by using metadata students are often unaware are being captured. Because students are unaware, RTE does not suffer from self-report and rater biases like many other types of measures (Kong, Wise, & Bhola, 2007; Rios, Liu, & Bridgeman, 2014). Beyond avoiding these forms of measurement bias, RTE can be easier to administer, score, and interpret than many self-report measures, advantages we describe more later.

3. The Goal of RTE is to Provide Students and Teachers with Immediate and Actionable Data on a Student's Self-management. The purpose of RTE in an SEL context is to use rapid guessing as an interim measure of self-management offered at multiple time points during the year to inform intervention and supplement other measures of self-management. These goals include two aspects: how RTE should be used as a measure, and how scores from that measure can be used to support effective intervention. We focus on the former in (3) and the latter in (4b). All of the potential uses (measurement and intervention) we discuss below should be supported with more validation research, some of which is already underway. We also recommend that decisions about students' self-management needs be based on RTE in conjunction with other measures. Invalid uses of RTE might involve using it as a sole measure to make determinations about interventions, especially if those determinations have consequences.

There are several ways we envision RTE being used as a measure. First, it can be utilized as part of a multiple-measures approach to assessing self-management. For example, RTE scores can be combined with formative assessments conducted informally by teachers over the course of the year, scores from more formal observation instruments, and scores from student surveys to identify students in need of self-management interventions. As an example of how this approach might look in practice, a district like SAUSD that administers a self-management survey in the spring could use RTE data obtained during fall and winter achievement test administrations to identify students who may have low self-management in advance of the spring survey administration. Further, using concurrent RTE and survey scores has several advantages, including safe-guarding against self-report bias. For instance, if a student reports high self-management but rapidly guesses often, then educators might worry about biased survey results, or at least use the discrepant data to foster conversation with the student.

Another potential measurement use of RTE is as an early warning indicator that a student might drop out, an outcome that is often driven in part by low academic motivation. Using behaviors that are manifestations of SEL constructs to predict drop-out is common in the early warning systems research (Allensworth & Easton, 2005; Balfanz & Boccanfuso, 2007). This literature identifies indicators that a student is likely to drop out in order to intervene early and get the student on track to graduate. Indicators include behaviors like course failures, suspensions/expulsions, and chronic absenteeism. As discussed earlier, our work shows a strong relationship between rapid guessing and these behaviors (Soland et al., 2017).

4(a). RTE is Easy to Use Because It Requires Little Specialized Knowledge Related to Administration, Scoring, or Interpretation. A major advantage of RTE is how easy it is to use in practice. Unlike many other measures, RTE requires practically no expertise to administer, score, and interpret. Further, whereas other measures can require users to wait for scores, RTE can be presented shortly after a test is completed. Assuming the district already administers a computer-based achievement test like MAP, measuring RTE does not require extra equipment, materials, or tools, which can reduce its cost. Even if a district does not already offer a computer-based achievement test, there are other opportunities to measure rapid guessing. For example, a manuscript in preparation by Soland, Wise, and Gao (2017) shows that rapid guessing occurs on surveys and tends to measure a similar construct, which means a survey itself can capture RTE data. This ease of use is one reason we describe RTE as a "benchmark" SEL measure: scores can be captured more frequently than when using surveys alone, which means educators have data between administrations of other measures.

4(b). The Simplicity of RTE Makes It Especially Useful in Promoting Self-management Supports for Students. In addition to its ease of use, RTE has other advantages that make it useful for educators trying to improve student self-management. For one, RTE is easy to interpret. A teacher can say that a hypothetical student with an RTE score of .85 rapidly guessed on 15% of the questions.[iii] Research further shows that RTE scores of below .90 are especially worrisome because the resultant subject test scores include so much rapid guessing, they may not be valid estimates of the student's achievement (DeMars & Wise, 2008). Our study (Soland et al., 2017) also shows that behaviors like low attendance are much higher for students with RTE values below .90 than for those students who did not rapidly guess (we hope to refine these thresholds so they are more specific to SEL-based interventions in future research). Students with RTE values below .90 may be good candidates for self-management interventions if they also show warning signs based on self-management surveys, teacher observations, or other data. There are a variety of effective self-management interventions, including personal goal-setting, self-monitoring, self-evaluation and recording, self-reinforcement, and self-charting (Briesch & Chafouleas, 2009). Students with RTE values below .90 and who exhibit other behaviors associated with academic disengagement like suspensions may also be candidates for drop-out prevention interventions, especially ones focused on academic motivation like those described by Balfanz, Herzog, and Mac Iver (2007).

5. RTE is Easily Scaled. Because RTE uses metadata already captured by many tests, it can be scaled quickly and easily. As a case in point, starting in the 2017-18 school year, RTE will be measured and reported for all students who use MAP assessments, which are administered to students in grades K-12, and are used across the U.S. in over 6,500 U.S. school systems. Over nine million students were assessed in math and reading across these systems during spring of 2016, totaling more than 12 million test events. Students will receive RTE data for all grades (K-12) and subjects tested. These effort metadata are automatically collected, with no need for schools to administer the assessments in different ways, or order a report at an additional cost to the MAP assessments themselves. Further, student RTE information from prior years will be made available so schools will be able to look at patterns of rapid guessing over time.

6. RTE Data Will be Reported Back to Educators and Students. Student-level RTE information will be measured and made available to educators, school leaders, and students in several ways.

Most importantly, RTE will be collected and reported on standard MAP student profile reports (which include, among other things, a student's test score, the standard error of measurement associated with the score, and normative information about the student's performance). These reports are available for review 24 hours after a student completes his or her testing, and will allow educators to quickly identify students who rapidly guessed. Additionally, the overall impact of a student's rapid-guessing behavior on his or her final achievement score will be measured and included in student reports. That is, MAP scores will be re-estimated based only on non-rapidly guessed item responses, which removes much of the bias from rapid guessing (Wise & Kingsbury, 2016). The difference in these adjusted and unadjusted scores indicates the extent to which student rapid-guessing behavior impacted a student's final test score. These data provide actionable information to students and educators by quantifying the impact of low self-management on achievement.

Assessment Description

7. RTE is Developmentally Appropriate for Grades 6-9. Evidence from our research indicates that RTE is a useful measure to track in middle school and early high school, a crucial transition period that often determines the likelihood that students will graduate (Mizelle & Irvin, 2000). Rates of rapid-guessing behavior increase as students get older, with 15% of students or higher in middle school and beyond showing levels of rapid guessing sufficient to potentially impact the validity of student scores (Soland, 2017; Wise, 2015). This general pattern in RTE is observed in our data nationwide. Research shows similar across-grade patterns in self-management, and low academic motivation more generally. For example, Balfanz, Herzog, and Mac Iver (2007) show that low academic motivation often begins in middle school and increases during the early high school years.

The types of questions students see on an achievement test also play a role in the appropriateness of RTE as a measure of rapid guessing. While response time metadata can be captured on any computer-based test, using particular types of computer-adaptive tests (CATs) like MAP can help ensure students are not rapidly guessing because items are far too easy or difficult for them. The CAT engine used by MAP selects items based on an estimate of that student's achievement that is re-evaluated after each item. Thus, students taking MAP should only receive items that are developmentally appropriate, and on material they have had an opportunity to learn. This facet of MAP means that RTE is not simply a proxy for academic ability (Wise & Kong, 2005) because students are rapidly guessing on items they have a reasonable probability of answering correctly.

8. Initial Evidence Suggests RTE is Culturally Appropriate. There are two aspects to ensuring RTE is a valid measure of rapid guessing behavior across racial, ethnic, linguistic, and cultural backgrounds. First, the achievement test from which response times are captured must be unbiased for these groups. While we cannot speak to the rigor of bias detection methods for other tests, MAP items undergo multiple sensitivity and fairness checks to ensure that all students are given equal opportunity to answer the item correctly based solely on their knowledge of the item content. Items are flagged and rewritten (or removed from the assessment altogether) if there is any evidence of cultural, linguistic, socio-economic, religious, gender, or geographic bias. Items that pass these initial tests are continually reviewed for the presence of differential item functioning (DIF), where students of the same ability level from different student groups of

interest are shown to have different probabilities of providing a correct answer to an item. Any items that are found to demonstrate even moderate DIF are subjected to additional reviews by content experts and, if necessary, removed from the assessment item bank.

Second, RTE must itself be unbiased across groups. Initial evidence suggests RTE is appropriate for students from a wide range of backgrounds. Soland et al. (2017) used a student sample from SAUSD, which has a high percentage of Hispanic, English-learner, and low-income students. The patterns of rapid guessing behavior in SAUSD are consistent with those in other districts and regions with different ethnic and socioeconomic compositions. For example, Soland (2017) finds consistent rates of rapid guessing for five different races across five different geographical regions in the U.S. Though we have not formally tested the measurement invariance of RTE across student subgroups, we intend to do so in the future.

9. RTE has High Potential to Help Students Become Better Learners by Making the Connection Between Self-management and Achievement Explicit. One way that RTE is unique as an SEL measure is that its impact on achievement can be made immediately apparent. Rapid-guessing behavior tends to bias observed test scores downwards, oftentimes by more than .25 standard deviations (Rios, Guo, Mao, & Liu, 2016). By re-scoring tests to account for rapid guessing, we can show students not only their RTE score, but also how much their achievement score *might have improved* if they had remained focused throughout the test. Making an explicit connection between self-management and test scores helps illuminate the complicated psychological processes that lead to low achievement and, thereby, provides more concrete opportunities to intervene. Students can see that low achievement is due not only to lack of content mastery, but also to the attitudes they hold about their abilities and the behaviors that result from those attitudes. If students observe that small changes in their self-management behaviors increase achievement, then there could be positive impacts on self-efficacy, the lack of which is oftentimes the root cause of poor self-management. The effect of directly showing students the connection between self-management and achievement is worthy of further study.

10. RTE is Supported by More than A Decade of Validation Evidence. As a measure of test motivation, RTE is supported by considerable validity evidence. The studies contributing to that evidence were cataloged by Wise (2015). Therefore, rather than describe that body of research in detail here, we instead provide a table from Wise (2015) listing those studies by type of evidence in Appendix 2. These studies tend to confirm RTE (1) demonstrates adequate levels of reliability, (2) is correlated with other measures of test motivation, (3) is not correlated with measures of academic ability, and (4) flags items as rapid guesses that yield scores that are correct at rates no better than chance. Research also shows that correlations between scores from measures of two related constructs tend to increase when rapid guesses from those measures are removed.

Despite the validity evidence supporting the use of RTE as a measure of test motivation, more work (beyond our initial study) needs to be done to validate the use of RTE as a measure of self-management. Much of this work is underway, and we allude to much of it throughout the submission, including the prototype. Additional studies should examine the goals of RTE as both a measure and intervention tool. In terms of RTE's purpose as a measure, our study only examined correlations between self-management and RTE in concurrent time periods. We plan to conduct a study examining how well RTE predicts later self-management survey scores, and

more distal behaviors/outcomes like graduation rates. As for RTE as a tool for intervention, work should be conducted with districts to examine the effect of giving teachers RTE as an interim benchmark on self-management, achievement, and behaviors like absenteeism.

Notes

---

[i] A copy of the manuscript, which is currently under peer review, is attached.

[ii] From a measurement perspective, self-management is a complicated construct. Rather than refer to a specific latent variable, it more frequently represents a collection of behaviors. For example, the survey used by Santa Ana Unified (and by all districts in the California Office to Reform Education or CORE) to measure self-management focuses entirely on whether students self-report exhibiting behaviors like coming to class prepared, following directions, and being able to work independently. Therefore, like other measures of self-management, rapid guessing is not exactly measuring a latent construct in the way that a growth mindset survey captures an unobservable belief that intelligence is malleable. Rather, rapid guessing is just one of several behaviors suggesting students may have trouble controlling their thoughts, behaviors, and emotions. To acknowledge this important if subtle distinction, we often refer to rapid guessing as a proxy for, rather than a measure of, self-management.

The distinction between a latent variable and observable behaviors also helps distinguish self-management from self-regulation. Though the two are highly related, and despite some disagreement in the relevant literature, in our proposal we think of self-regulation as a latent variable measuring how well a student maintains control over thoughts, emotions, and actions, and self-management primarily as whether students take those actions associated with self-regulation. That is, low self-management is often the observable manifestation of low self-regulation in the form of actions and behaviors associated with the latent trait.

[iii] One complication in using and interpreting RTE, however, is that teachers will need to be somewhat careful in how they describe the result. If students know exactly how responses are flagged as motivated or unmotivated, then the measure can be gamed, one admitted disadvantage. Therefore, the measure will likely be more useful if flagged items are described as "unmotivated" or "disengaged" rather than mention rapid guessing specifically. This concern is one reason why our reports describe RTE-based scores as the "Percent of Disengaged Responses".